Technical Appendix on Sparse Bayesian Regression

Loic Le Folgoc, Hervé Delingette, Antonio Criminisi, Nicholas Ayache

Technical Appendix on Sparse Bayesian Regression

Loïc Le Folgoc[1], Hervé Delingette[1], Antonio Criminisi[2], and Nicholas Ayache[1]

[1] Asclepios Research Project, INRIA Sophia Antipolis, France
[2] Machine Learning and Perception Group, Microsoft Research Cambridge, UK

Abstract. We report the technical details for a sparse bayesian approach to regression. It can be seen as an extension of the Relevance Vector Machine of Tipping *et al* [1] to a more general setting that can handle vector-valued regression and generic quadratic priors.

1 Quadratic Energies & Marginal Likelihood of the Data

We want to minimize in $\boldsymbol{w}$ the following cost:

$$E(\boldsymbol{w}) = (\boldsymbol{t} - \Phi\boldsymbol{w})^\intercal \beta \mathbf{H} (\boldsymbol{t} - \Phi\boldsymbol{w}) + \boldsymbol{w}^\intercal (\mathbf{A} + \lambda \mathbf{R}) \boldsymbol{w} \tag{1}$$

and jointly optimize in $\mathbf{A} = \mathrm{diag}(\mathbf{A}_i)$, β, λ. $\boldsymbol{w}|\boldsymbol{t}, \mathbf{A}, \lambda, \beta$ follows a Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$, where

$$\boldsymbol{\mu} = \Sigma \Phi^\intercal \beta \mathbf{H} \boldsymbol{t}, \quad \Sigma = (\mathbf{A} + \lambda \mathbf{R} + \Phi^\intercal \beta \mathbf{H} \Phi)^{-1} \tag{2}$$

A key element is that the distribution of $\boldsymbol{t}|\mathbf{A}$ is also Gaussian, $\boldsymbol{t}|\mathbf{A} \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$, with

$$\mathbf{C} = (\beta \mathbf{H})^{-1} + \Phi (\mathbf{A} + \lambda \mathbf{R})^{-1} \Phi^\intercal \tag{3}$$

Indeed, we see that:

$$\begin{aligned} p(\boldsymbol{t}|\mathbf{A}) &= \int p(\boldsymbol{t}|\boldsymbol{w}) p(\boldsymbol{w}|\mathbf{A}) d\boldsymbol{w} \\ &\propto \int \exp -\frac{1}{2} (\boldsymbol{t} - \Phi\boldsymbol{w})^\intercal \beta \mathbf{H} (\boldsymbol{t} - \Phi\boldsymbol{w}) \cdot \exp -\frac{1}{2} \boldsymbol{w}^\intercal (\mathbf{A} + \lambda \mathbf{R}) \boldsymbol{w} \cdot d\boldsymbol{w} \\ &\propto \exp -\frac{1}{2} \boldsymbol{\mu}^\intercal \Sigma^{-1} \boldsymbol{\mu} \cdot \exp -\frac{1}{2} \boldsymbol{t}^\intercal \beta \mathbf{H} \boldsymbol{t} \cdot \int \exp -\frac{1}{2} (\boldsymbol{w} - \boldsymbol{\mu})^\intercal \Sigma^{-1} (\boldsymbol{w} - \boldsymbol{\mu}) d\boldsymbol{w} \end{aligned}$$

where the integral sums to 1. Thus using (2),

$$p(\boldsymbol{t}|\mathbf{A}) \propto \exp -\frac{1}{2} \boldsymbol{t}^\intercal (\beta \mathbf{H}) \boldsymbol{t} - \frac{1}{2} \boldsymbol{t}^\intercal (\beta \mathbf{H}) \Phi \Sigma \Phi^\intercal (\beta \mathbf{H})) \boldsymbol{t}$$

$\boldsymbol{t}|\mathbf{A}$ is Gaussian since the distribution is proportional to a Gaussian, and by identification it ensues that $\mathbf{C}^{-1} = \beta \mathbf{H} - (\beta \mathbf{H}) \Phi \Sigma \Phi^\intercal (\beta \mathbf{H})$. We then get the desired result using a matrix inversion identity. The fast RVM algorithm proceeds by iteratively implementing a single change to one of the $\mathbf{A}_i$'s on the block-diagonal matrix $\mathbf{A}$; specifically the one that maximizes the increase of a quantity known as the *evidence*, $\log p(\boldsymbol{t}|\mathbf{A})$, then re-estimating the parameters of the conditional posterior $\boldsymbol{w}|\boldsymbol{t}, \mathbf{A}, \lambda, \beta$ using (2). The algorithm starts with all $\mathbf{A}_i$ set to ∞. The computations involve rank-one "block" updates; it also turns out that the optimal $\mathbf{A}_i$'s are rank one matrices (so we actually have rank-one updates).

2 Computation of the gain in evidence for a given action

We want to evaluate the change in $\log p(\boldsymbol{t}|\mathbf{A}_{-i}, \mathbf{A}_i)$ when a single prior weight $\mathbf{A}_i$ is changed. Recall that $\boldsymbol{t}|\mathbf{A}_{-i}, \mathbf{A}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{C})$, thus

$$\log p(\boldsymbol{t}|\mathbf{A}_{-i}, \mathbf{A}_i) = -1/2 \left\{\log |\mathbf{C}| + \boldsymbol{t}^\intercal \mathbf{C}^{-1} \boldsymbol{t}\right\} \tag{4}$$

Now let's single out the contribution of $\mathbf{A}_i$, for each of the two terms above. Noting the form of eq. (3), let $\mathbf{L} = (\mathbf{A} + \lambda \mathbf{R})^{-1}$. We can single out the contribution of the ith basis to $\mathbf{L}$ first, as a rank one update:

$$\mathbf{L} = \begin{pmatrix} \mathbf{L}_{-i} & 0 \\ 0 & 0 \end{pmatrix} + \mathrm{U}_i \, l_i \, \mathrm{U}_i^\intercal \tag{5}$$

with $\mathrm{U}_i^\intercal = \left((\lambda \mathbf{L}_{-i} \mathbf{R}_i)^\intercal \ \mathrm{I}^\intercal\right)$ and $l_i = \{\mathbf{A}_i + \lambda \mathbf{R}_{ii} - (\lambda \mathbf{R}_i)^\intercal \mathbf{L}_{-i} (\lambda \mathbf{R}_i)\}^{-1} \triangleq \{\mathbf{A}_i + \kappa_i\}^{-1}$. Note also that any basis for which $\mathbf{A}_j = \infty$ can be disregarded, since its corresponding line and column in $\mathbf{L}$ and $\mathbf{L}_{-i}$ will be null. We can interpret $\mathbf{L}$ as a square matrix of dimension the number of active bases (including the basis under consideration), and the algorithmic complexity of matrix operations involving $\mathbf{L}$ or $\mathbf{L}_{-i}$ will indeed be related to the reduced sized of these matrices. This is also true for U_i, as a direct consequence, and for all of the other quantities involved.

Injecting (5) into (3) gives a decomposition of $\mathbf{C}$ into the sum of a term that does not depend on the ith basis and of a rank-one term:

$$\mathbf{C} = \mathbf{C}_{-i} + \Phi^i \mathrm{U}_i \, l_i \, \mathrm{U}_i^\intercal \Phi^{i\intercal} \tag{6}$$

Φ is superscripted with i to recall that along with all the other active bases, the ith basis ϕ_i is present in this matrix. Using rank-one updates for, respectively, the determinant and the inverse, and letting $\mathbf{C}_{-i}^{-1} \triangleq (\mathbf{C}_{-i})^{-1}$, we get the two following expressions:

$$|\mathbf{C}| = |\mathbf{C}_{-i}| \, |l_i| \, |l_i^{-1} + \mathrm{U}_i^\intercal \Phi^{i\intercal} \mathbf{C}_{-i}^{-1} \Phi^i \mathrm{U}_i| \tag{7}$$

$$\boldsymbol{t}^\intercal \mathbf{C}^{-1} \boldsymbol{t} = \boldsymbol{t}^\intercal \left(\mathbf{C}_{-i}^{-1} - \mathbf{C}_{-i}^{-1} \Phi^i \mathrm{U}_i \left\{l_i^{-1} + \mathrm{U}_i^\intercal \Phi^{i\intercal} \mathbf{C}_{-i}^{-1} \Phi^i \mathrm{U}_i\right\}^{-1} \mathrm{U}_i^\intercal \Phi^{i\intercal} \mathbf{C}_{-i}^{-1}\right) \boldsymbol{t} \tag{8}$$

These quantities rewrite more compactly if we introduce appropriate notations. Namely, let $q_j(i) \triangleq \phi_j^\intercal \mathbf{C}_{-i}^{-1} \boldsymbol{t} \in \mathbb{R}^d$ and $s_{jk}(i) \triangleq \phi_j^\intercal \mathbf{C}_{-i}^{-1} \phi_k \in \mathcal{M}_{d,d}$. The concatenation of these q_j's for all active bases plus the basis under scrutiny (total of m bases), a.k.a $\boldsymbol{q}_i \in \mathbb{R}^{d \times m}$, will come in helpful. Similarly $\boldsymbol{s}_i \in \mathcal{M}_{d\times m, d\times m}$ will denote the matrix with $s_{kl}(i)$ as (k,l)th coefficient, where indices span the set of active bases plus the ith basis. Now, let $q^i \triangleq \mathrm{U}_i^\intercal \Phi^{i\intercal} \mathbf{C}_{-i}^{-1} \boldsymbol{t} = \mathrm{U}_i^\intercal \boldsymbol{q}_i \in \mathbb{R}^d$, and $s^i \triangleq \mathrm{U}_i^\intercal \Phi^{i\intercal} \mathbf{C}_{-i}^{-1} \Phi^i \mathrm{U}_i = \mathrm{U}_i^\intercal \boldsymbol{s}_i \mathrm{U}_i \in \mathcal{M}_{d,d}$. With these notations in hand and recalling that $l_i^{-1} = \mathbf{A}_i + \kappa_i$, we can rewrite eq. (7) and eq. (8) as:

$$\log |\mathbf{C}| = \log |\mathbf{C}_{-i}| - \log |\mathbf{A}_i + \kappa_i| + \log |\mathbf{A}_i + \kappa_i + s^i| \tag{9}$$

$$\boldsymbol{t}^\intercal \mathbf{C}^{-1} \boldsymbol{t} = \boldsymbol{t}^\intercal \mathbf{C}_{-i}^{-1} \boldsymbol{t} - q^{i\intercal} \left\{\mathbf{A}_i + \kappa_i + s^i\right\}^{-1} q^i \tag{10}$$

Ignoring the terms that do not depend on the ith basis, we see that the contribution to the evidence of the model for a given value of $\mathbf{A}_i$ is directly related to:

$$l(\mathbf{A}_i) = \log|\mathbf{A}_i + \kappa_i| - \log|\mathbf{A}_i + \kappa_i + s^i| + q^{i^\intercal}\left\{\mathbf{A}_i + \kappa_i + s^i\right\}^{-1} q^i \tag{11}$$

Naturally if $\lambda = 0$ (no additional regularization) this comes down to the regular RVM, with $q^i = q_i$, $s^i = s_{ii}$ and $\kappa_i = 0$.

3 Maximization of the gain in evidence

If $q^i q^{i^\intercal} - s^i$ has no positive eigenvalue, the maximum $\mathbf{A}_i$ lies at infinity and the basis should remain inactive, or be removed. Otherwise the gradient of Eq. (11) provides ground to look for rank-one maximizers $\mathbf{A}_i = \alpha_i \eta_i \eta_i^\intercal$. To that aim we compute $n_i = s^{i^{-1}} q^i$ and

$$a_i = \frac{(n_i^\intercal s^i n_i)^2}{(n_i^\intercal q^i)^2 - n_i^\intercal s^i n_i} - n_i^\intercal \kappa_i n_i \tag{12}$$

If $a_i \geq 0$ the maximizer is given by $\alpha_i = a_i$ and $\eta_i = n_i$. Otherwise ($a_i < 0$) we set $\alpha_i = 0$ and numerically solve over the optimal orientation η_i. This latter case arises when the regularization level alone is sufficient to make additional ”shrinkage” unnecessary.

4 Update of $\boldsymbol{\lambda}$

We derive an update rule via an expectation-maximization procedure. Knowing $\boldsymbol{w}$, it would be straightforward to derive an estimate of λ by maximizing the log-likelihood of $\boldsymbol{w}$ or the posterior of λ given $\boldsymbol{w}$. However $\boldsymbol{w}$ is hidden in our model. Instead, we try to maximize the log-likelihood on average (i.e. to minimize the average loss): $\max_\lambda \mathbb{E}_{\boldsymbol{w}\sim\mathcal{N}(\boldsymbol{\mu},\Sigma)}[\log p(\boldsymbol{t},\boldsymbol{w}|\mathbf{A},\lambda,\beta)|\mathbf{A}^*,\beta^*]$, where $\mathbf{A}^*$ and β^* are our current estimates of the respective quantities. Discarding terms which are constants of λ, we obtain the following:

$$\lambda^* = \arg\max_\lambda -\frac{\lambda}{2}\mathrm{tr}(\boldsymbol{\Sigma}\mathbf{R}) + \frac{1}{2}\log|\mathbf{A} + \lambda\mathbf{R}| - \frac{\lambda}{2}\boldsymbol{\mu}^\intercal\mathbf{R}\boldsymbol{\mu} \tag{13}$$

Deriving leads to:

$$\frac{\partial}{\partial\lambda} f(\lambda) \propto \mathrm{tr}(\{\mathbf{A} + \lambda\mathbf{R}\}^{-1}\mathbf{R}) - \mathrm{tr}(\boldsymbol{\Sigma}\mathbf{R}) - \boldsymbol{\mu}^\intercal\mathbf{R}\boldsymbol{\mu} \tag{14}$$

This is a decreasing function of λ and thus has at most one zero. If $\frac{\partial}{\partial\lambda}f(\lambda)$ is negative at the origin, $\lambda^* = 0$. Otherwise, we optimize by using the Newton method on a log scale. This is motivated by the fact that the function of interest is not only decreasing, but also smooth and convex. Lastly note that $\{\mathbf{A} + \lambda\mathbf{R}\}^{-1}\mathbf{R} = \{\mathbf{R}^{-1}\mathbf{A} + \lambda\mathrm{I}\}^{-1}$, so we can compute the eigenvalues δ_k of $\mathbf{A}^{1/2}\mathbf{R}^{-1}\mathbf{A}^{1/2}$ once and rely on the fact that $\mathrm{tr}(\{\mathbf{A} + \lambda\mathbf{R}\}^{-1}\mathbf{R}) = \sum_k 1/(\delta_k + \lambda)$ to avoid repeated matrix inversions.

References

1. Tipping, M.E., Faul, A.C., et al.: Fast marginal likelihood maximisation for sparse bayesian models. In: Workshop on artificial intelligence and statistics. Volume 1., Jan (2003)